Are Contrastive Explanations Useful? *

James Forrest[1], Somayajulu Sripada[1], Wei Pang[2], and George M. Coghill[1]

[1] University of Aberdeen {j.forrest , yaji.sripada , g.coghill} @abdn.ac.uk
[2] Heriot-Watt University W.Pang@hw.ac.uk

Abstract. From the user perspective (data subjects and data controllers), useful explanations of ML decisions are selective, contrastive and social. In this paper, we describe an algorithm for generating selective and contrastive explanations and experimentally study its usefulness to users.

Keywords: Interpretable ML · Contrastive Explanations · XAI.

1 Introduction

Machine Learning (ML) models are making ever increasing numbers of decisions that impact people's lives. This makes it important to have explanations that enable those who develop and deploy ML models and those who are subject to their decisions to examine these models to ensure that they are effective and fair. Miller 2017 proposes three desiderata for explanations - explanations should be selective, contrastive and social. Explanations should be contrastive, explaining how other events could have happened rather than explaining the event that did happen. Explanations should be selective, only presenting information relevant to the recipient. Explanations should be social because they are a communication between explainer and recipient and need to respect the recipient's needs.

Because operationally deployed ML models could come from an extensive taxonomy of ML models (from decision trees to deep neural networks) an equally large taxonomy of interpretation methods have been developed (Guidotti et al., 2018). Not all these interpretation methods fulfil Miller's desiderata. Wachter et al. 2017 propose that Contrastive Explanations best fulfil these criteria. In a Contrastive Explanation, the recipient is shown counterfactuals, where the decision model would make different decisions, as the explanation. This work is a test of some claims made for Contrastive Explanations.

In order to have data professionals make better use of explanations, Kaur et al 2020 suggest that ML explanations have improved use of HCI to aid understanding in users. Especially to promote better alignment of the users' mental models and the conceptual models of the Interpretable Machine Learning (IML) tool and help the users move from reflexive to deliberative thinking.

* Supported by EPSRC DTP Grant Number EP/N509814/1

In this paper, we present a Contrastive Explanation technique built on opt-aiNet, an optimisation algorithm based on an Immune Inspired Algorithm (Timmis and Edmonds, 2010) that fulfils Miller's desiderata. We evaluate our Contrastive Explanation technique in comparison to the contrastive tool DiCE (Mothilal et al., 2020), as well as Feature Saliency explanations.

2 Explanation Generation

A Contrastive Explanation is a collection of counterfactuals; each counterfactual shows the changes to the original event (fact) that would produce the wanted event (foil). These counterfactuals are most effective when selective; the more selective the explanation, the fewer causes the explanation contains, the simpler and more comprehensible the explanation. The counterfactuals should produce foil events that are as near to fact event as possible to be more actionable to the recipient

2.1 opt-aiNet

Our counterfactual generation algorithm one of a family of algorithms called Artificial Immune Systems (AIS), which use the Immune System as a metaphor to solve problems. The counterfactuals are generated by the AIS algorithm opt-aiNet, an existing algorithm that we have applied to the domain of explanation generation.

The features of an opt-aiNet system (illustrated in fig 1) are:

- The counterfactuals in this AIS take the cell in the Immune System as their metaphor.
- The counterfactuals (cells) clone themselves with mutations, a mutation changes the explanation randomly.
- The distance between a cell and the target is the affinity; cells near the target have high affinity, and cells far from the target have low affinity. Mutation varies inversely with the affinity; high-affinity cells mutate little while low-affinity cells mutate much more.
- Counterfactuals (cells) interact with each other with high-affinity counterfactuals suppressing similar counterfactuals with lower affinity. This removal of similar counterfactuals helps enforce the diversity of counterfactuals.

Pseudocode for opt-aiNet (Timmis and Edmonds, 2010) adapted for explanation counterfactuals

1: Produce initial population of counterfactuals (cells)
2: **repeat**
3: **repeat**
4: Generate N clone counterfactuals for each counterfactual. All must be classified as foil class
5: Mutate each counterfactual in inverse proportion to the affinity of the parent counterfactual

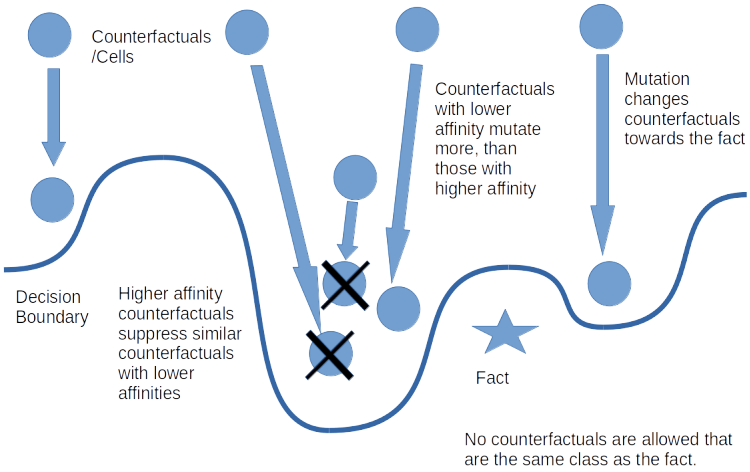

Fig. 1. Illustration of opt-aiNet for Counterfactual Generation

Determine the affinity of all counterfactuals
Keep counterfactual with the highest affinity suppress all others
until average affinity lower than previous iteration
Determine highest affinity counterfactuals, suppress similar lower affinity counterfactuals
Introduce a proportion of new randomly generated counterfactuals
until stopping condition met

The algorithm is based on the the Ruby programming code provided by Brownlee 2012, which was then reimplemented by us in Python.

2.2 DiCE

DiCE is an IML tool that produces counterfactual explanations produced in (Mothilal et al., 2020). Optimising on the proximity of the explanation to the fact, this algorithm also optimises to maximise the diversity of the set of explanations it creates.

3 Experiments

Our experiments evaluated the explanations of a model's decision. The evaluation methodology followed that suggested in Hoffman et al. 2018 for Explainable AI (XAI) and Interpretable Machine Learning (IML) evaluations. We used a Test of Performance, in which the participant's comprehension of the explanation of the model's decision was determined by a proxy task of reversing an adverse

decision of the model. We used a Test of Satisfaction where the participants answered questions on their satisfaction with the explanation. We did not use a Test of Comprehension, which evaluates the participant's mental models, because these tests are very hard to execute well and are hard to use one test across different types of explanation, whereas our Test of Performance could be used for all the explanation types.

The data used in this experiment comes from the lending dataset maintained by Kaggle at https://www.kaggle.com/wordsforthewise/lending-club, which holds data about credit decisions. The decision model used TensorFlow to create a Deep Neural Network.

The task simulated a situation where a data subject received a negative prediction and needed to know the changes to their record that would give a positive prediction. A negative prediction was used in this experiment because data subjects are presumed to care about negative credit predictions more than positive credit predictions. And gives a natural task of amending the record to get a positive result.

Participants were recruited using Amazon Mechanical Turk. The tests were conducted between explanation types, with each experiment using one explanation only. Every participant was shown the same three records with only the explanations changing between participants.

The Test of Performance was to change the values of the record so that the result changed from a refusal of credit where the model scored the record < 0.5 to an acceptance where the model scores ≥ 0.5. The participants were asked to change the record by the smallest amount which would cause the model to classify it positively. Those participants with the best understanding of the model should produce a score ≥ 0.5 but close to 0.5.

The Test of Satisfaction was six questions on the user's satisfaction with the explanation rated on a five-point Likert scale.

The questions were:

1. I understand this explanation of how the decision was made.
2. This explanation of how the decision was made is satisfying.
3. The explanation of how the decision was made has sufficient detail.
4. This explanation of how the decision was made contains irrelevant details.
5. The record is useful to the goal of having the loan application accepted
6. This explanation lets me judge the trustworthiness of the explanation.

Further evaluation was done using two Feature Saliency explanations and a control of no explanation. Feature Saliency explanations for tabular data produce a set of pairs of input Features with their Importance. The tools used to generate these Feature Saliency explanations were LIME (Ribeiro et al., 2016) and SHAP (Lundberg and Lee, 2017), two very popular tools in IML explanation. There were two presentations used of the Feature Saliency explanation type: a bar-chart and a Natural Language Generation (NLG) text. Both presentations have a table showing the feature/importance pairs. With an additional element of either the same data presented as a bar-chart, or a text description. The participants who received No Explanation as a control only saw the data in the record.

4 Results

4.1 Test of Performance

The Test of Performance was conducted on whether the participant's alterations to the record successfully produced a change to the models decision from declined to accepted (Correct) or not (Incorrect), and evaluated using the Chi-Squared test. The results are shown in Table 1. Contrastive Explanations show no significant difference in whether opt-aiNet or DiCE generate the counterfactuals or whether one or three counterfactuals were used in the explanation. There is no significant difference in performance when the best performing Contrastive explanantion (opt-aiNet with three counterfactuals) is compared to other types of explanations. Comparing the explanation to 'No Explanation' where the participants are only shown the values in the record as a control explanation shows that this has the same performance as any given explanation. There are no significant differences in performance because this is preliminary work with small numbers of participants. Future work might involve doing these tests of explanations with greater numbers of participants to achieve significant results.

Explanations	Total	Correct	Incorrect	% Correct	significance from χ^2 test with opt-aiNet with 3cf
opt-aiNet with one counterfactual	30	19	11	63%	1
DiCE with one counterfactual	27	16	11	59%	0.59
opt-aiNet with three counterfactuals	30	20	10	67%	NA
DiCE with three counterfactuals	30	19	11	63%	1
Feature Saliency with bar-chart	63	41	22	65%	0.934
NLG with table	60	38	22	63%	0.938
No Explanation	60	40	20	67%	1

Table 1. Comparison on Test of Performance of different explanations

4.2 Test of Satisfaction

The results for the Test of Satisfaction are shown in Table 2, which shows the mean values for the six five-point Likert scale questions. The participants had higher satisfaction ratings for Contrastive Explanations with three counterfactuals rather than one counterfactual. The satisfaction ratings are similar for both counterfactual generation tools. The opt-aiNet with three counterfactuals has similar satisfaction ratings to the Feature Saliency explanations with either

Explanation Type	Q 1	Q 2	Q 3	Q 4	Q 5	Q 6
opt-aiNet with one counterfactual	3.93	3.59 *	3.67	3.33	3.77	3.47
DiCE with one counterfactual	3.66	3.36 *	3.55	3.48	3.82	3.33
opt-aiNet with three counterfactuals	3.93	4.27	3.80	3.07	4.13	3.87
DiCE with three counterfactuals	4.09	4.34	4.00	3.17	4.17	4.09
Feature Saliency with bar-chart	3.95	4.00	4.06	3.45	3.95	4.12
NLG with table	4.10	4.08	4.07	3.45	3.95	4.12
No Explanation	3.63	3.47 *	3.57	3.22	3.90	3.85

Table 2. Mean values of the five point likert questions in the Test of Satisfaction

** Significant differences between explanations and opt-AINet with three counterfactuals*

bar-chart or NLG. Further, the control 'No Explanation' has lower satisfaction ratings than these three explanations.

The significance of the results was evaluated using Mann-Whitney, compared to the opt-aiNet with three counterfactuals. The significant differences were for the second question, 'This explanation of how the decision was made is satisfying', where No Explanation and Contrastive Explanations with one counterfactual had lower ratings.

5 Discussion

The Test of Performance produced two unexpected results. Firstly, that Contrastive Explanation with one counterfactual performed unexpectedly poorly compared to other explanations. Because if the record is changed according to the values in the counterfactual, then the classification will change. Secondly, the participants who received the 'No Explanation' of just the values in the record performed as well at the task as those who received an explanation, even though their satisfaction ratings were lower.

Examining why this might be, figure 2 shows box-plots of the decision models scores of the participant's records. The experiment instructions asked participants to make the smallest changes to the record that would give a positive classification. For the Contrastive Explanations the box-plots show the decision model scores are close to the decision boundary at 0.5, but for the other explanations the average score is further from the decision boundary and have a larger range of scores. This shows the Contrastive Explanations produce more optimal results, whereas, Feature Saliency explanations produce explanations that are still effective at changing the classification, but do so in a less optimal way (the score is further from the decision boundary).

The 'No Explanation' is effective at changing the classification but poor at producing an optimal result, as the average score is the furthest from the decision boundary. The 'No Explanation' is showing the participants own mental models of the lending domain without being altered by IML explanations. Most participants who saw the 'No Explanation' have a good Mental Model about how to change a record to get credit approval e.g. have a high income and good credit grade. But both participants who saw a Feature Saliency (bar-chart and NLG) explanation and 'No explanation' did not know how much to change the record by to change the record by to get a different classification as this information is not given in the explanations.

Fig. 2. Box-plots of decision model's scores of participants altered records, for explanations

6 Conclusion

This preliminary work shows that the opt-aiNet algorithm produces counterfactuals as good as the established algorithm DiCE. Contrastive Explanations of the model are not necessarily more effective than Feature Saliency explanations or just showing the participants the data in the record used by the ML model. Contrastive Explanations do enable the participants to make more optimal changes to the record, that are nearer to the decision boundary. Participants have their own mental models of how well known domains, this allows them to be effective at changing the decisions of classifiers without use of IML explanations. Future work will require more data and experiments to discover what the effects of explanations of ML decisions are on users.

Bibliography

J. Brownlee. *Clever Algorithms*. 2012. ISBN 9781446785065. URL http://www.cleveralgorithms.com.

R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, D. Pedreschi, and F. Giannotti. A Survey Of Methods For Explaining Black Box Models. 51(5), 2018. ISSN 03600300. https://doi.org/10.1145/3236009. URL http://arxiv.org/abs/1802.01933.

R. R. Hoffman, S. T. Mueller, G. Klein, and J. Litman. Metrics for Explainable AI: Challenges and Prospects. pages 1–50, 2018. URL https://arxiv.org/abs/1812.04608.

H. Kaur, H. Nori, S. Jenkins, R. Caruana, H. Wallach, and J. Wortman Vaughan. Interpreting Interpretability: Understanding Data Scientists' Use of Interpretability Tools for Machine Learning. *CHI Conference on Human Factors in Computing Systems Proceedings*, pages 1–14, 2020. https://doi.org/10.1145/3313831.3376219.

S. Lundberg and S.-I. Lee. A Unified Approach to Interpreting Model Predictions. 2017. ISSN 10495258. URL https://arxiv.org/abs/1705.07874.

T. Miller. Explanation in Artificial Intelligence: Insights from the Social Sciences. 2017. https://doi.org/arXiv:1706.07269v1. URL http://arxiv.org/abs/1706.07269.

R. K. Mothilal, A. Sharma, and C. Tan. Explaining machine learning classifiers through diverse counterfactual explanations. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 607–617, New York, NY, USA, 1 2020. ACM. ISBN 9781450369367. https://doi.org/10.1145/3351095.3372850. URL http://arxiv.org/abs/1905.07697http://dl.acm.org/doi/10.1145/3351095.3372850.

M. T. Ribeiro, S. Singh, and C. Guestrin. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. 2016. ISSN 9781450321389. https://doi.org/10.1145/1235. URL http://arxiv.org/abs/1602.04938.

J. Timmis and C. Edmonds. A Comment on Opt-AiNET: An Immune Network Algorithm for Optimisation. (May):308–317, 2010. https://doi.org/10.1007/978-3-540-24854-5_32.

S. Wachter, B. Mittelstadt, and L. Floridi. Why a right to explanation of automated decision-making does not exist in the General Data Protection Regulation. *International Data Privacy Law*, pages 1–47, 2017. URL http://arxiv.org/abs/1606.08813.